Classification automatique dans les sous-espaces discriminants de Fisher

Charles Bouveyron[1], Camille Brunet[2]

1- SAMOS-MATISSE, CES, UMR CNRS 8174 – Université Paris 1
2- IBISC TADIB, FRE CNRS 3190 – Université d'Evry Val d'Essonne

Résumé : Nous considérons dans ce travail le problème du *clustering* en grande dimension. Nous proposons de modéliser les données par un modèle de mélange gaussien dans un sous-espace discriminant de dimension inférieure à la dimension de l'espace original. Nous proposons pour ce faire un algorithme d'estimation appelé Fisher-EM. Cette approche améliore les performances de classification et permet une représentation visuelle de l'agencement des données en grande dimension.

Abstract : This work focuses on the problem of high-dimensional clustering. The data are modeled by a Gaussian mixture model in a discriminative subspace with an intrinsic dimension lower than the dimension of the original space. An estimation algorithm, called Fisher-EM, is proposed. This approach outperforms existing methods and provides a visual representation of the high-dimensional data.

1 Introduction

Une partie des méthodes de classification souffre du bien connu *fléau de la dimension* et voit leur performance décroître à mesure que la dimension des observations augmente. En effet, les données de grande dimension présentent deux particularités qui rendent difficile l'exploitation et l'extraction de l'information qu'elles contiennent : d'une part les données de grande dimension vivent dans des espaces de dimension intrinsèque plus petite que la dimension de l'espace original et, d'autre part, ces données peuvent être bruitées. A cela s'ajoutent une difficulté de représentation graphique et de visualisation de ces données complexes. Étant donné que la dimension de l'espace des données observées est le plus souvent bien supérieure à leur dimension intrinsèque, il est théoriquement possible de réduire la dimension de l'espace initial sans perdre d'information. C'est dans cette optique que classiquement, dans un problème de classification automatique sur des données de grande dimension, certains prétraitements basés sur des extractions de caractéristiques, comme l'Analyse en Composante Principale (ACP), ou de sélection de variables sont effectués. Cependant, ces méthodes de réduction de dimension ne prennent pas en compte l'objectif de classification, particulièrement dans le cadre de la classification non supervisée. En effet, une des rares approches qui combine la réduction de dimension et la classification se trouve dans le cadre de la classification supervisée : il s'agit de l'analyse discriminante de Fisher. Cette méthode cherche une représentation des données dans un nouvel espace qui permet de discriminer au mieux les classes. L'analyse discriminante de Fisher (chap. 18

de [1]) consiste dans un premier temps à projeter les données de l'espace initial sur des axes discriminants qui maximisent le rapport de la variance interclasse sur la variance intra classe, puis à classer les données dans cette nouvelle représentation.

Nous proposons dans ce travail d'adapter les procédures traditionnelles de classification automatique par modèles de mélange pour que les étapes de modélisation et de classification soient réalisées dans des sous-espaces discriminants. L'objectif de notre approche est triple : d'une part, améliorer les techniques de classification non supervisée en recherchant un sous-espace discriminant, d'autre part, éviter des problèmes d'estimation des paramètres liés à la grande dimension et enfin permettre une représentation visuelle de l'agencement des données en grande dimension.

2 *Clustering* dans les sous-espaces discriminants

L'idée de notre approche repose sur le fait que les données observées sont une transformation des "vraies" données (non observées) de dimension intrinsèque faible et sur le fait qu'un sous-espace de dimension $K-1$ est théoriquement suffisant pour discriminer K classes. Nous proposons donc de modéliser et de classer les données dans un sous-espace discriminant de dimension $d = K-1$. Cette stratégie permettra d'obtenir à la fois un regroupement optimal, un modèle parcimonieux et une visualisation discriminante des données.

2.1 Le modèle

Le problème de la classification automatique réside dans la recherche d'une partition des données sans la connaissance *a priori* du nombre de classes qu'il faut trouver. Un moyen traditionnel de modéliser les données est d'utiliser un modèle de mélange qui fait l'hypothèse que chaque groupe peut être caractérisé par une distribution de probabilités. Dans notre étude nous nous référons au cas paramétrique où les densités de chaque classe sont modélisées par des lois gaussiennes. Nous faisons donc l'hypothèse que les données non observées $\{x_1, \ldots, x_n\}$ constituent un échantillon de n réalisations indépendantes du vecteur aléatoire X à valeur dans $\mathbb{R}^d$ et que les données observées $\{y_1, \ldots, y_n\}$ à valeur dans $\mathbb{R}^p$ sont une transformation linéaire des $\{x_1, \ldots, x_n\}$:

$$y = x\ V + \epsilon,$$

où $x \in \mathbb{R}^d$, $y \in \mathbb{R}^p$, $V \in \mathcal{M}_{d\times p}$ et $\epsilon \in \mathbb{R}^p$ est un terme de bruit. Nous supposons en outre que la densité de X s'écrit de la façon suivante :

$$f(x) = \sum_{k=1}^{K} \pi_k \phi(x, \theta_k),$$

où K est le nombre total de groupes, ϕ est la densité de la loi normale de paramètres $\theta_k = (\mu_k, \Sigma_k)$ et π_k est la proportion de la $k$ème composante du mélange. Les paramètres du modèle du mélange sont traditionnellement estimés par la méthode du maximum de vraisemblance au moyen de l'algorithme itératif *Expectation-Maximization* (EM) [2]. Dans notre approche, la modélisation étant faite dans le sous-espace de dimension d et non dans l'espace des données observées, il est nécessaire d'adjoindre à l'algorithme EM une étape d'estimation de la transformation V.

2.2 Procédure d'estimation : l'algorithme Fisher-EM

Du fait de la nature du modèle proposé précédemment, la procédure d'estimation basée sur l'algorithme EM comportera trois étapes : tout d'abord, une *étape E* qui calcule les probabilités *a posteriori* d'appartenir aux différents groupes dans le sous-espace de dimension d, une *étape F* qui calcule la transformation V conditionnellement aux probabilités *a posteriori* obtenues dans l'étape E et qui permet le passage de l'espace de Fisher ($\mathbb{R}^d$) vers l'espace des observations ($\mathbb{R}^p$), et enfin une *étape M* qui estime les paramètres du mélange décrit dans le sous-espace de dimension d par maximisation de la vraisemblance. Cette procédure d'estimation itérative, appelée algorithme *Fisher-EM* par la suite, s'écrit de la manière suivante :

Étape E : Cette étape calcule, à l'itération (q), les probabilités *a posteriori* t_{ik} d'appartenir à la $k$ème composante du mélange (dans le sous-espace de dimension d) :

$$t_{ik}^{(q)} = \frac{\pi_k^{(q)} \phi(x_i, \theta_k^{(q)})}{\sum_{k=1}^{K} \pi_k^{(q)} \phi(x_i, \theta_k^{(q)})},$$

où ϕ représente la densité gaussienne de moyenne μ_k et de matrice de covariance Σ_k dans l'espace discriminant de dimension $d = K - 1$.

Étape F : A l'itération (q), cette étape estime la matrice de transformation V par le calcul des axes discriminants, dits de Fisher, conditionnellement aux probabilités *a posteriori* obtenues à l'*étape E*. Les axes discriminants sont les axes qui maximisent le rapport entre la variance interclasse et la variance intra classe. Grâce au théorème de Huyghens, les axes de projection recherchés satisfont le problème d'optimisation suivant :

$$max_u \frac{u^t B^{(q)} u}{u^t S u},$$

où $S = \frac{1}{n} \sum_{i=1}^{n} (y_i - m)^t (y_i - m)$ et $m = \frac{1}{n} \sum_{i=1}^{n} y_i$, sont respectivement la matrice de

covariance et la moyenne du nuage de points observés, et :

$$B^{(q)}(t_{ik}) = \frac{1}{n}\sum_{k=1}^{K} n_k^{(q)}(m_k^{(q)} - m)^t(m_k^{(q)} - m),$$
$$n_k^{(q)} = \sum_{i=1}^{N} t_{ik}^{(q)},\ m_k^{(q)} = \frac{1}{N}\sum_{i=1}^{N} t_{ik}^{(q)} y_i\ .$$

Les solutions de ce problème d'optimisation sont les vecteurs propres de la matrice $S^{-1}B^{(q)}$.

Étape M : Cette étape estime par maximisation de la vraisemblance les paramètres des K composantes dans l'espace des axes discriminants de Fisher : les probabilités *a priori* π_k, les moyennes μ_k et les d termes diagonaux σ_{kr} ($r \in \{1, \ldots, d\}$) de la matrice de covariance Σ_k diagonale par construction puisque les d composantes retenues dans l'espace de Fisher sont orthogonales entre elles :

$$\mu_k^{(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)} x_i^{(q)}}{\sum_{i=1}^{N} t_{ik}^{(q)}},\ \pi_k^{(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)}}{N}$$
$$\sigma_{kr}^{(q+1)} = \frac{\sum_{i=1}^{N} t_{ik}^{(q)}(x_{i(r)}^{(q)} - \mu_{k(r)})^2}{\sum_{i=1}^{N} t_{ik}^{(q)}},$$

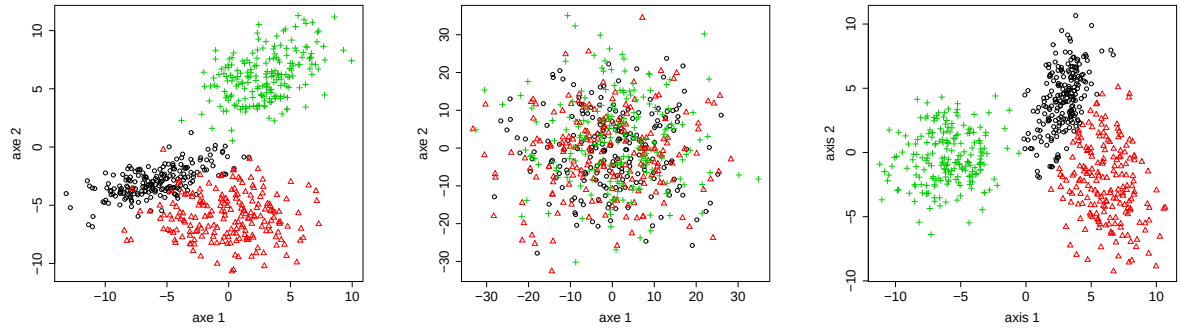
avec $x_i^{(q)} = y\, U^{(q)}$ où $U^{(q)}$ est la matrice de transformation inverse de taille $p \times d$ contenant les d plus grands vecteurs propres de $S^{-1}B^{(q)}$. Classiquement, une telle approche sur des données de grande dimension présente des problèmes d'estimation des paramètres du mélange. Or, l'approche exposée dans ce paragraphe offre l'avantage de ne nécessiter que le calcul d'un nombre réduit de paramètres puisque leur estimation est faite dans un sous-espace de dimension $d = K - 1$ plus petit que la dimension de l'espace original et indépendant de la dimension de l'espace original. De plus, les problèmes classiques d'inversement et d'estimation des matrices de covariance sont évités.

3 Résultats expérimentaux

Nous proposons dans cette section d'appliquer l'algorithme Fisher-EM à un jeu de données simulées afin de mettre en évidence les deux principaux avantages de notre approche : la visualisation et la performance en grande dimension.

3.1 Visualisation dans le sous-espace discriminant

Afin de vérifier les performances de l'approche proposée au paragraphe précédent, sur des données de grande dimension et bruitées, nous avons simulé 600 observations dans



(a) Données dans l'espace latent. (b) Projection sur les 2 axes de l'ACP. (c) Sous-espace discriminant estimé.

FIG. 1 – Jeu de données de 600 observations réparties en 3 classes

$\mathbb{R}^{25}$ composées de trois classes de dimensions intrinsèques 2, auxquelles nous avons ajouté vingt-trois dimensions de bruit gaussien $\mathcal{N}(0, 10)$. La figure 1(a) présente les données simulées dans leur espace latent (de dimension intrinsèque 2) tandis que la figure 1(b) montre la projection des données observées (de dimension 25) sur les 2 premières composantes principales de l'ACP. Il apparaît nettement que cette dernière représentation ne permet pas de distinguer les trois groupes, alors qu'il existe par construction deux dimensions intrinsèques qui séparent totalement les classes. La figure 1(c) présente le résultat de l'algorithme Fisher-EM. Nous pouvons observer que l'estimation des deux axes discriminants par l'algorithme Fisher-EM permet d'obtenir à la fois un regroupement de très bonne qualité (96% correct) et une visualisation explicative des données.

3.2 Influence de la dimension

Nous allons à présent comparer les performances de l'algorithme Fisher-EM avec des méthodes standards de *clustering* telles que l'algorithme EM, l'algorithme EM réalisé sur les composantes principales de l'ACP (ACP+EM) et les *k-means*. Comme précédemment, les données simulées appartiennent à 3 groupes, chacun modélisé par une densité gaussienne de dimension 2 auxquelles nous avons ajouté de une à quinze dimensions supplémentaires de bruit gaussien suivant la loi $\mathcal{N}(0, 10)$. Les mêmes conditions d'initialisation sont appliquées pour chaque méthode. Nous avons choisi ici une initialisation par les paramètres du mélange telle qu'elle est proposée par McLachlan et Peel [3]. La figure 2 illustre l'influence de la dimension sur les performances de classification obtenues par l'algorithme Fisher-EM, EM, EM+ACP et *k-means*. Nous remarquons tout d'abord que la dimension n'a pas une influence significative sur les performances de Fisher-EM. Le taux de classification correct moyen est de l'ordre de 0.90 quelle que soit la dimension. D'autre part, nous remarquons que les algorithmes EM et ACP+EM voient leur performance décroître presque linéairement avec l'augmentation de la dimension. Enfin, l'algorithme des *k-means* présente une performance peu satisfaisante dans l'espace initial qui se dégrade

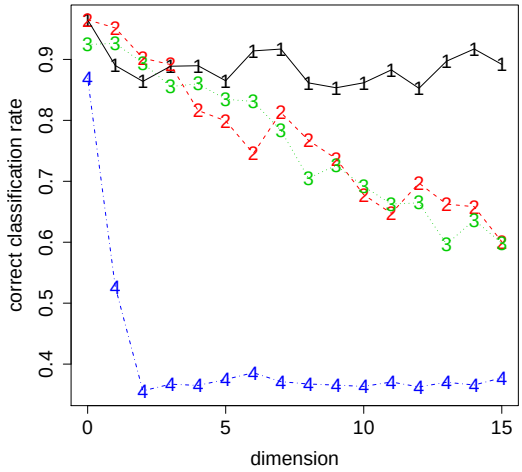


FIG. 2 – Évolution du taux de classification correct selon la dimension et selon l'approche réalisée : Fisher-EM (1-noir), EM (2-rouge), ACP+EM (3-vert), *k-means* (4-bleu).

très rapidement avec l'augmentation de la dimension. Nous pouvons noter en revanche, que comme tout algorithme de type EM, Fisher-EM présente une légère instabilité due à l'étape d'initialisation. On observe toutefois que le taux de classification correct varie entre 0.87 et 0.96, ce qui reste toujours satisfaisant.

4 Conclusion

Dans ce travail, nous avons proposé une méthode de classification automatique dans les sous-espaces discriminants de Fisher basée sur le modèle de mélange. Un algorithme d'estimation, appelé Fisher-EM, basé sur l'algorithme EM est proposé pour l'estimation des sous-espaces discriminants et des paramètres du mélange. Cette approche s'est avérée plus performante que les méthodes traditionnelles de classification automatique sur simulation. Elle s'est révélée être peu sensible à l'augmentation de la dimension et aux données bruitées. De plus, le nombre de paramètres à estimer est indépendant de la dimension originale des données. Enfin, cette approche permet une visualisation explicite de l'agencement des données en grande dimension.

Références

[1] G. Saporta. *Probabilités, analyse des données et statistique*. Editions Technip, 2006.

[2] A. Dempster, N. Laird, and D. Robin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1) :1–38, 1977.

[3] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Interscience, New York, 2000.